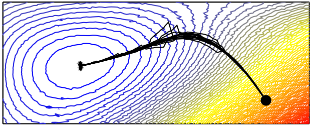
Differentiating electromagnetic showers in a sampling calorimeter

Max Aehle, University of Kaiserslautern-Landau
Mihály Novák, CERN
Vassil Vassilev, Princeton
Nicolas R. Gauger, University of Kaiserslautern-Landau
Lukas Heinrich, TU Munich
Michael Kagan, SLAC
David Lange, Princeton

CHEP '24, Krakow
October 19th–25th, 2024

arxiv.org/abs/2405.07944



Outline:

1. Mathematical Challenges
2. AD for G4HepEm/HepEmShow
3. Future Perspectives

Algorithmic Differentiation (AD) / Diff. Programming

- Set of techniques to **evaluate derivatives of computer-implemented functions**.
- Useful e. g. for **gradient-based optimization** of a computationally heavy loss function, finding optimal design parameters, model parameters etc.
- **Forward mode** of AD with a single AD input x: For each number a handled by the primal program, keep track of $\dot{a} = \frac{\partial a}{\partial x}$, augmenting all real-arithmetic operations:

$$c = a + b \rightsquigarrow \dot{c} = \dot{a} + \dot{b}$$

$$c = a \cdot b \rightsquigarrow \dot{c} = \dot{a} \cdot b + a \cdot \dot{b}$$

  etc. Run-time and memory performance asymptotics match those of numerical differentiation, but AD is exact.
- **Reverse mode of AD**: More complicated. Allows to compute a gradient of a single AD output w. r. t. many AD inputs in one stroke. $\rightsquigarrow$ Extremely useful for optimization.

AD Tools

AD tools identify real-arithmetic operations in the primal program and insert the appropriate AD logic. Different mechanisms have been reported:

- Source Transformation
- Operator Overloading, e. g. CoDiPack from RPTU.
- Hooking into the compiler, e. g. Clad by V. Vassilev.
- Dynamic binary instrumentation of machine code, Derivgrind from RPTU (M. Aehle's PhD topic).
- Hardware

Video (7 min) about Derivgrind+LibreOffice Calc:
https://t1p.de/tt4ne

Can AD be useful for HEP?

- There were (gradient-free) optimization studies already while designing the LHC.[1]
- Studies concerned with single experiments have shown ample room for optimization.[2]
- There is a collaboration[3] and a yearly workshop[4] on bringing AD into several fields of fundamental physics.
- No differentiated version of Geant4 yet.
- To make it work, we need to solve **technical and mathematical challenges**.

---

[1] S. Russenschuck, T. Tortschanoff. Mathematical Optimization of Superconducting Accelerator Magnets, IEEE Trans. on Magnetics 30 (5) 1994.

[2] T. Dorigo, Geometry optimization of a muon-electron scattering detector, Physics Open 4 (2020) 100022

[3] MODE Collaboration, coordinated by **T. Dorigo**, `mode-collaboration.github.io`

[4] `https://indico.cern.ch/event/1242538/`

AD of Geant4 – Many Challenges

Technical

- Geant4 is really big ($\sim$ 1 M lines of C++ code). AD tools *do not require rewriting*, but usually still *some manual efforts*. $\rightsquigarrow$ Machine-code-based AD with Derivgrind solves this aspect.
- Performance degrades, memory required to store the *tape*, etc.

Mathematical

The derivatives we compute with AD may not be the ones we need.
$\rightsquigarrow$ This talk: Study AD for Geant4-like particle simulation.

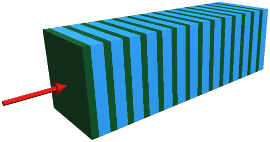
G4HepEm and HepEmShow

G4HepEm R&D

- R & D project by Mihaly Novak, Jonas Hahnfeld, Ben Morgan
- Simulation of electromagnetic showers (e^-, e^+, γ)
- Isolates the required data and functionalities from Geant4, well documented, customizable.

compact EM shower simulation
HepEmShow

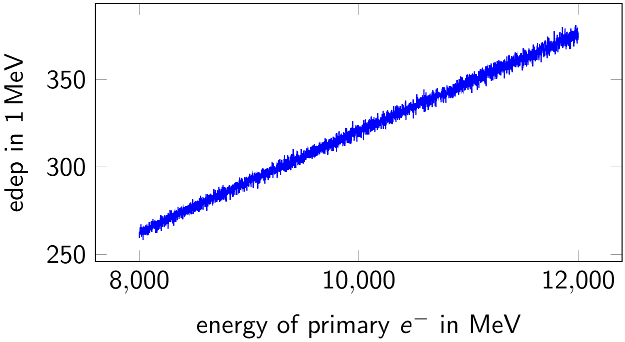
- Created by Mihaly Novak
- Self-contained application using G4HepEm but not Geant4.
- Simulates electromagnetic showers in a sandwich calorimeter.



github.com/mnovak42/g4hepem
github.com/mnovak42/hepemshow

Mathematical Challenges

Average energy deposition as a function of primary energy:

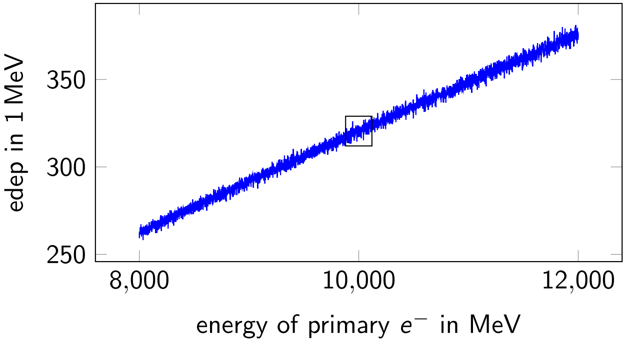


Energy deposition of an EM shower in some layer of a sampling calorimeter, depending on the energy of the primary particle. Simulated with G4HepEm/HepEmShow, 1k events per data point.

Even though the "noise" has a low magnitude, its derivative can have a large magnitude! The derivative can also be zero with probability 1...

Mathematical Challenges

Average energy deposition as a function of primary energy:

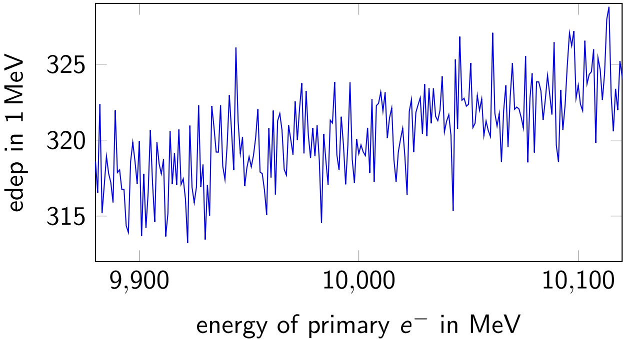


Energy deposition of an EM shower in some layer of a sampling calorimeter, depending on the energy of the primary particle. Simulated with G4HepEm/HepEmShow, 1k events per data point.

Even though the "noise" has a low magnitude, its derivative can have a large magnitude! The derivative can also be zero with probability 1...

Mathematical Challenges

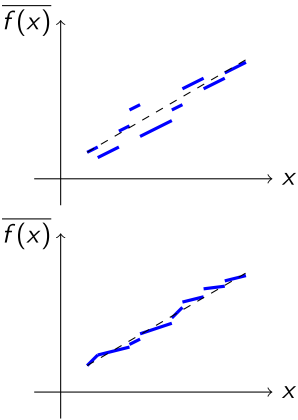
Average energy deposition as a function of primary energy:



Energy deposition of an EM shower in some layer of a sampling calorimeter, depending on the energy of the primary particle. Simulated with G4HepEm/HepEmShow, 1k events per data point.

Even though the "noise" has a low magnitude, its derivative can have a large magnitude! The derivative can also be zero with probability 1...

Which one of these?

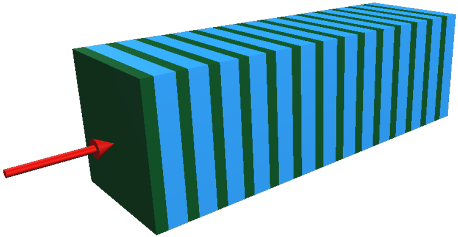


Mathematical Challenges

- Is the variance of the derivative sufficiently low? Otherwise we'd have to average over too many events to get a trustworthy result...
- When we average the derivative where it exists, do we get the derivative of the averages?
- Can the AD derivatives be useful for optimization?

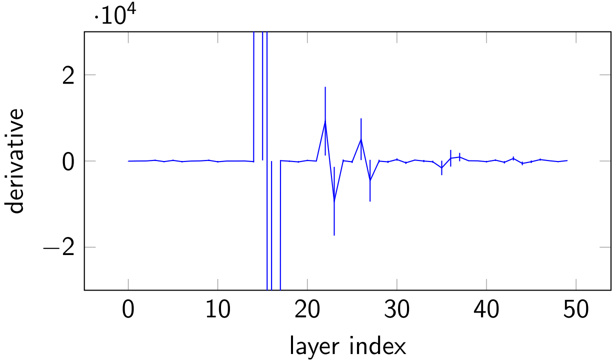
$\frac{\partial\,(\text{Energy Deposition})}{\partial\,(\text{Primary Particle Energy})}$ for the fully detailed simulation

Applied the AD tool CoDiPack (took $\sim$3 days).



$\frac{\partial\,(\text{Energy Deposition})}{\partial\,(\text{Primary Particle Energy})}$ for the fully detailed simulation

Applied the AD tool CoDiPack (took ∼3 days).

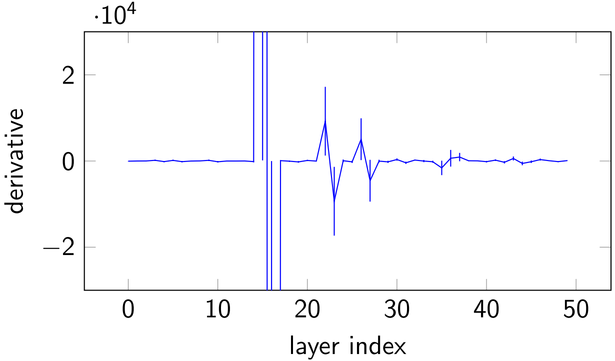


AD at 10 GeV

Look at the vertical axis range. These noisy derivatives can hardly be useful.

$\frac{\partial\,(\text{Energy Deposition})}{\partial\,(\text{Primary Particle Energy})}$ for the fully detailed simulation

Applied the AD tool CoDiPack (took ∼3 days).

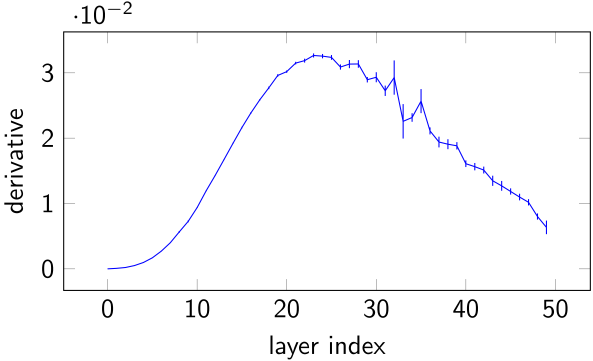


AD at 10 GeV

Look at the vertical axis range. These noisy derivatives can hardly be useful.

⇝ **Deactivate multiple scattering** in the simulation.

$\frac{\partial\,(\text{Energy Deposition})}{\partial\,(\text{Primary Particle Energy})}$ with simplifications

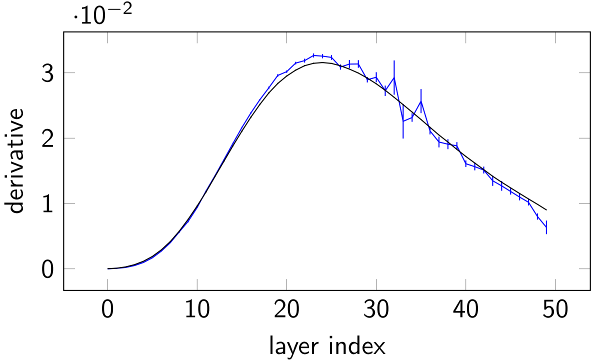


AD at 10 GeV

multiple scattering disabled

This looks much better!

$\frac{\partial\,(\text{Energy Deposition})}{\partial\,(\text{Primary Particle Energy})}$ with simplifications

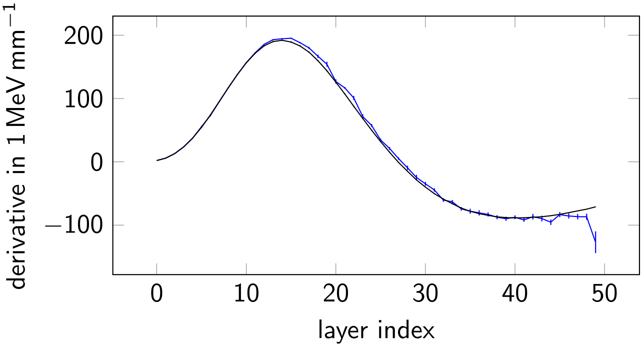


AD at 10 GeV
Diff.quot. at 9.9...10.1 GeV
multiple scattering disabled

This looks much better!

What about other AD inputs?

$\frac{\partial\,(\text{Energy Deposition})}{\partial\,(\text{Absorber Thickness})}$ with simplifications



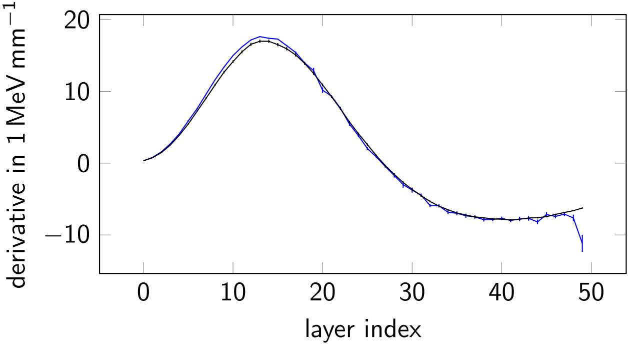
AD at 2.3 mm

Diff.quot. at 2.29. . . 2.31 mm

multiple scattering and fluctuation disabled

Good agreement as well.

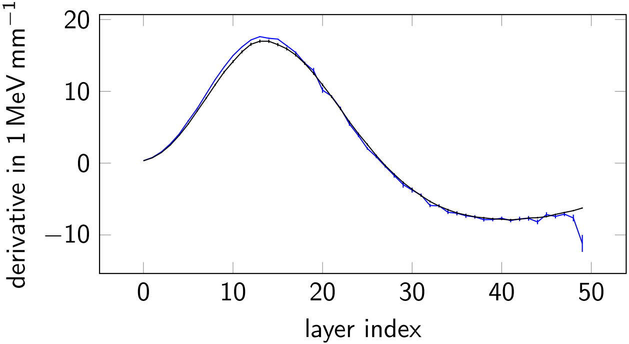
$\frac{\partial\,(\text{Energy Deposition})}{\partial\,(\text{Gap Thickness})}$ with simplifications



AD at 5.7 mm
Diff.quot. at 5.65...5.75 mm
multiple scattering and fluctuation disabled

Good agreement as well.

$\frac{\partial\,(\text{Energy Deposition})}{\partial\,(\text{Gap Thickness})}$ with simplifications

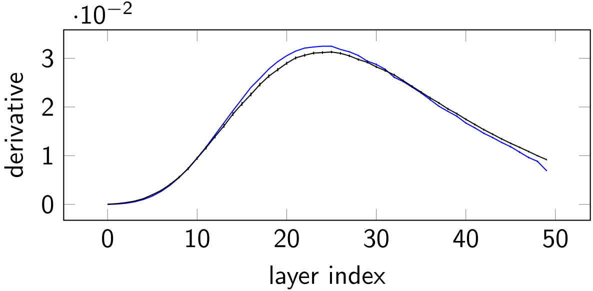


AD at 5.7 mm
Diff.quot. at 5.65. . . 5.75 mm
multiple scattering and fluctuation disabled

Good agreement as well.

Going back to the $\frac{\partial\,(\text{Energy Deposition})}{\partial\,(\text{Primary Particle Energy})}$ plot: Let's reduce the step size of the difference quotient to reduce the truncation error. Let's take more events to make the error bars smaller.

$\frac{\partial\,(\text{Energy Deposition})}{\partial\,(\text{Primary Particle Energy})}$ with simplifications, many events



AD at 10 GeV
Diff.quot. at 9.995...10.005 GeV
multiple scattering and fluctuation disabled

The ∼5 % deviation is **small but statistically significant**.

Hypothesis: The 5 % bias has to do with this ⟶
There is a novel method[5] to handle some of these constructs; implementation will be lots of work.

```
double p = /* diff'able */;
if( rng->flat() < p ){
  // ... do something ...
}
```

[5] G. Arya et al. Automatic Differentiation of Programs with Discrete Randomness. NeurIPS 2022.

Gradient-Based Optimization Problem

Given a

- primary energy e and
- absorber thickness a,

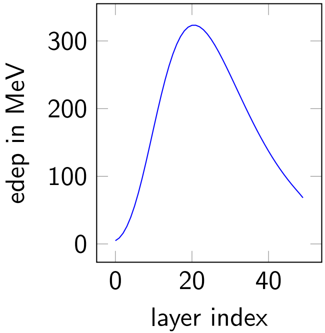
HepEmShow computes the resulting edep distribution $f_i(e, a)$ in the layers $i = 0, \ldots, 49$.

Given $f(e^*, a^*)$ (e. g. measured), can we infer e^*, a^*?

$\rightsquigarrow$ For the loss function

$$L(e, a) = \|f(e, a) - f(e^*, a^*)\|_2^2,$$

find $\min_{e,a} L(e, a)$!



Gradient-Based Optimization Setup

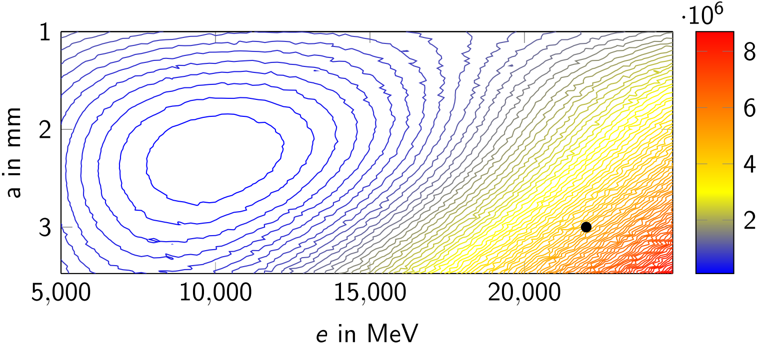
$$\underbrace{\frac{\partial L}{\partial(e,a)}(e_i,a_i)}_{2\times 1} = \underbrace{\frac{\partial L}{\partial f}\left(f(e_i,a_i) \right)}_{1\times 50} \underbrace{\cdot \frac{\partial f}{\partial(e,a)}(e_i,a_i)}_{50\times 2}$$

Starting with some (e_0, a_0), iteratively,

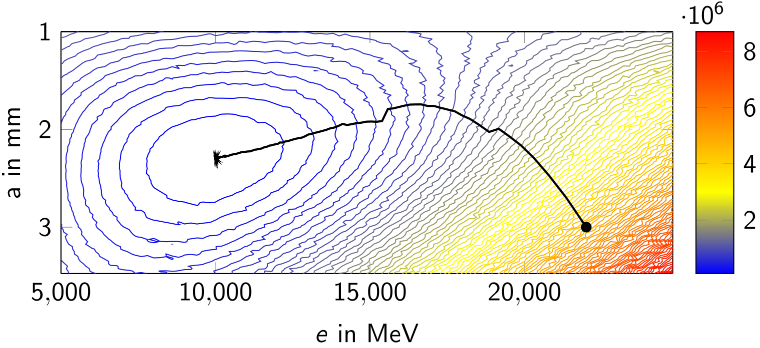
1. Evaluate $f(e_i, a_i)$, i. e. run HepEmShow without AD,
2. Evaluate $\frac{\partial L}{\partial f}(f(e_i,a_i)) = 2(f(e_i,a_i) - f(e^*,a^*))^T$,
3. Evaluate the vector-jacobian product (vjp) with $\frac{\partial f}{\partial(e,a)}(e,a)$, i. e. run HepEmShow with reverse-mode AD.
4. Gradient descent step with step-sizes λ_e, λ_a:

$$e_{i+1} = e_i - \lambda_e \cdot \frac{\partial L}{\partial e}(e_i,a_i)$$
$$a_{i+1} = a_i - \lambda_a \cdot \frac{\partial L}{\partial a}(e_i,a_i)$$

Gradient-Based Optimization Results

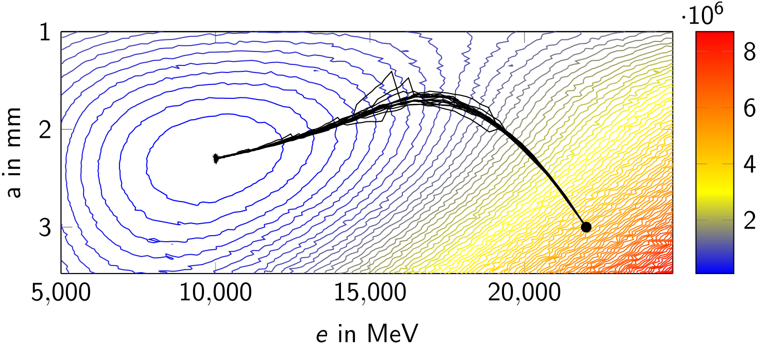


Gradient-Based Optimization Results



350 gradient descent steps with $\lambda_e = 1$, $\lambda_a = 10^{-7}\,\mathrm{mm}^2\,\mathrm{MeV}^{-2}$, 1000 events per iteration. Starting from $e_0 = 22\,000\,\mathrm{MeV}$, $a_0 = 3\,\mathrm{mm}$, converging to the minimizer $e^* = 10\,000\,\mathrm{MeV}$, $a^* = 2.3\,\mathrm{mm}$.

Gradient-Based Optimization Results



350 gradient descent steps with $\lambda_e = 1$, $\lambda_a = 10^{-7}\,\mathrm{mm}^2\,\mathrm{MeV}^{-2}$, 1000 events per iteration. Starting from $e_0 = 22\,000\,\mathrm{MeV}$, $a_0 = 3\,\mathrm{mm}$, converging to the minimizer $e^* = 10\,000\,\mathrm{MeV}$, $a^* = 2.3\,\mathrm{mm}$.

Paths are stochastic, of course.

Mathematical Challenges

- Is the variance of the derivative sufficiently low? Otherwise we'd have to average over too many events to get a trustworthy result...
- When we average the derivative where it exists, do we get the derivative of the averages?
- Can the AD derivatives be useful for optimization?

Mathematical Challenges

- Is the variance of the derivative sufficiently low? Otherwise we'd have to average over too many events to get a trustworthy result...
  When disabling multiple scattering, the variance is OK.
- When we average the derivative where it exists, do we get the derivative of the averages?
  Not exactly, but up to 5 %, which is fine for optimization.
- Can the AD derivatives be useful for optimization?
  Yes!

Summary

Integrating AD capabilities in HEP detector simulations works in principle, and can be worthwhile

- from an application perspective: to understand how design parameters affect objective functions, and enable efficient gradient-based optimization; probably there are more applications than that;
- from a AD research perspective: To our knowledge, no tooling for unbiased AD of generic Monte-Carlo codes available so far.

The community should always keep **alternative approaches** in mind:

- numerical differentiation
- gradient-free optimization
- differentiable surrogate models

Next Steps

- Find out what the problem with MSC is.
- Scale this work from G4HepEm/HepEmShow to Geant4.
- Use the derivatives for scientifically challenging questions such as detector design.

Contact Information

Max Aehle, max.aehle@scicomp.uni-kl.de,

Nicolas R. Gauger, nicolas.gauger@scicomp.uni-kl.de

Scientific Computing Group
University of Kaiserslautern-Landau (RPTU)

Joint work with Mihály Novák, Vassil Vassilev, Lukas Heinrich, Michael Kagan and David Lange.

Optimization Using Pathwise Algorithmic Derivatives of Electromagnetic Shower Simulations
http://arxiv.org/abs/2405.07944